Detecion of opinion Spam in online reviews

Shaswat Rungta

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India

Detection of opinion spam in online reviews

B.Tech project submission report submitted in partial fulfillment of the requirements for the degree of

Bachelors of Technology

in

Computer Science and Engineering

by

Shaswat Rungta
(Roll: 111CS0156)

with the supervision of

Prof. Banshidhar Majhi
NIT Rourkela

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India
May 2015

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India.

Date : .....

Certificate

This is to certify that the work in the thesis entitled *Detection of opinion spam in online reviews* by Shaswat Rungta is a record of an original research work carried out under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Banshidhar Majhi
Professor
Department of CSE, NIT Rourkela

Acknowledgment

This dissertation, though an individual work, has beneted in various ways from several people. Whilst it would be simple to name them all, it would not be easy to thank them enough. The enthusiastic guidance and support of Prof. B. Majhi, who inspired me to stretch beyond my limits. His profound insight has guided my thinking to improve the nal product.

Many thanks to my fellow research colleagues. It gives me a sense of happiness to be with you all. Special thanks to Smriti Singh, who provided help and support whenever I needed it.

Finally, I am grateful to all my friends for continuous motivation and encouragement. Last but not the least to my family having faith in me and always supporting me.

Shaswat Rungta

Abstract

The rise of Internet has led to consumers constantly and increasingly review and research products and services online. Consequently, websites that garner such reviews become primary targets for opinion Spam, which essentially means to sway public opinion by posting deceptive reviews. In this work, we have worked on integrating linguistic features and N-gram modeling to develop a feature set that can be used to detect authentic sounding yet fake reviews. A data set of 1600 reviews from 20 different hotels [7, 8] is used for experimentation and results. From the findings, we also try to figure out what can possibly be the factors that help to detect the spammers, and, additionally, make suggestions that can be incorporated by websites to control Spam based on user information.

Keywordsopinion spam, spam detection, hotel reviews

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Men are social animals and this feature is exhibited profoundly when we are buying commodities for ourselves, It is human nature to spend judiciously and try to get the best out of any investment. We also , inevitably, find ourselves in a state of ignorance about the things we intend to buy, may it be in terms of the commodity's specification, usage, performance, or whether it is a suitable buy for us. The only out for us then is to ask someone who knows what we do not, and based on how much trust we have on the person, we take a purchase decision. Here a commodity can be materialistic item like a mobile phone or a service like massage. This model leaves quite interesting manifestations in terms of forming public opinion of the commodity, as described in the next sections

1.1 History

Before e-commerce became mainstream, people bought stuff from off-line outlets. This meant before buying anything, in case of a dilemma they could take suggestions from people around them, friends, family or acquaintances. For example, if someone wanted to buy a mobile phone, he would ask a few of his known ones, *who may or may not* have used the product, about possible buying options. *Based on these suggestions*, he would them go to an off-line store to buy the phone he has decided. He may take into account the shopkeeper's opinion and then buy or reject the phone.(The flow chart is shown in 1.1)

As more online commerce spaces starting popping up, this flow chart got altered

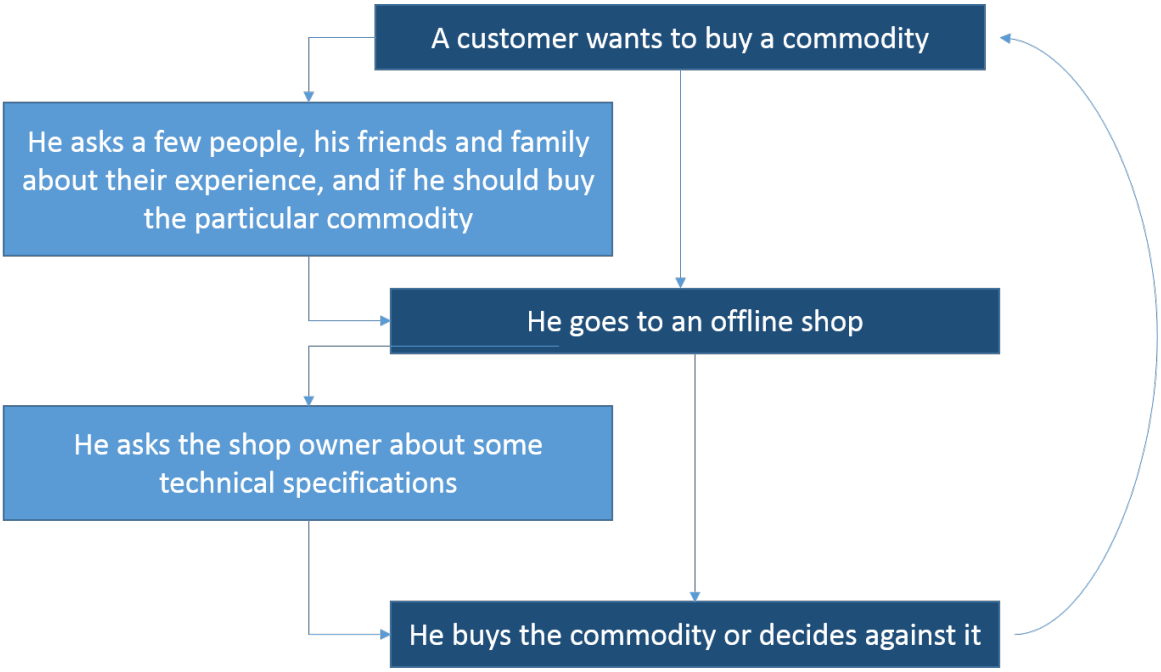

Figure 1.1: Traditional Customer Purchase Flowchart

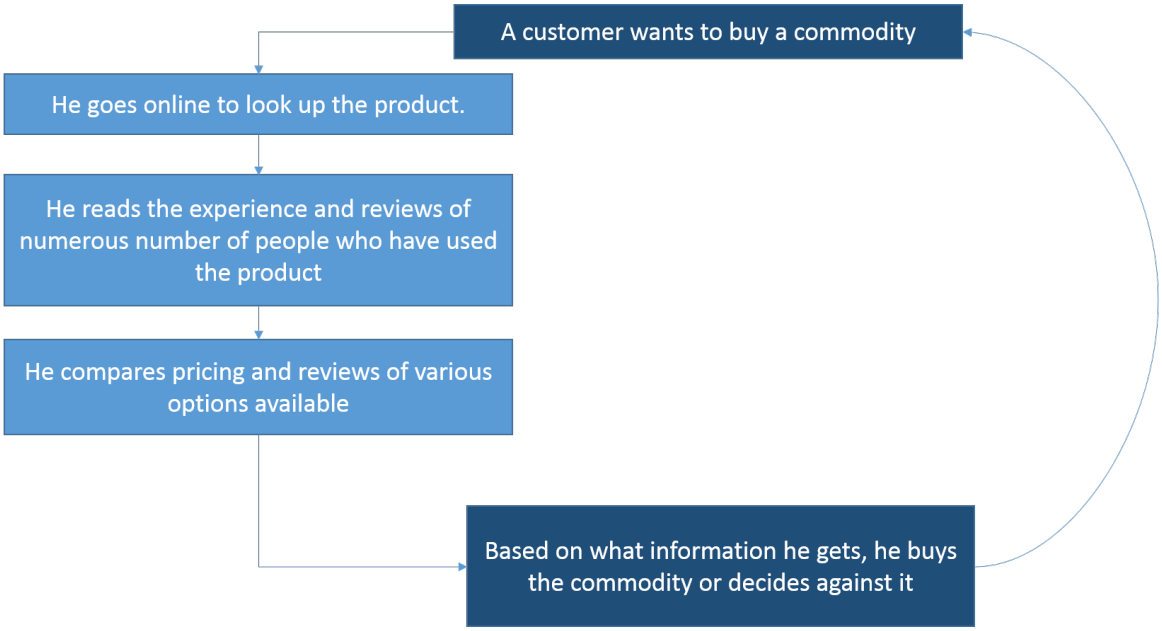

Figure 1.2: Technology-oriented Customer Purchase Flowchart

a bit, as shown in 1.2. These sites gave people who bought items an option to review the items, profoundly giving them a tool to express their agreement or dissent on the quality of the product or service. A new customer can now visit these sites and read a number of reviews before deciding to buy into the product. However a major change from the previous model is that these reviews are from people one barely knows. In other words, reviews from complete strangers. In the transition from the traditional setup to an on-line one, people began taking "suggestions" and reviews from a much larger community of people who apparently knew about the product. While the product's factual specification could be checked from various sources, the authenticity of the review people had written about the product was largely undetermined. Nevertheless, one's opinion of the product was heavily swayed by what reviews they read.

Given that the world we live in is not Utopian, this change in shopping pattern led to interesting avenues to manipulate and misguide what the customer choose to purchase.

1.2 Social and Economic Impact

In the light of the upsurge of the on-line shopping, people have started trusting on-line reviews more and more. In a community driven environment, if the prominent populace says that a service is bad, it is usually presumed to be bad. As the opinions garner over time, such seeded opinion can make or break public opinion about products. With enough resources and proper planning, a few average products can be portrayed as way better than they are. On the flip side, such resources can also be used to bring down somebody else's product.

As people buy products after reading the reviews, and what people buy affects the seller selling the product, by transitivity, the kind of reviews that a product attracts is of concern to the sellers, be it on-line or off-line. This means that a positive review on product would bring in sales and a negative one would reduce them. Thus this economics forces the sellers to invest time and resources to get their reviews straight, either ethically or otherwise.

Also given the competition that exists in current market space for all commodities, howsoever large-scale or small, it becomes an ardent need to stay ahead of the curve and be better in the public eye than one's competitors. This may mean portraying one's own product as better or, at times, portraying someone-else's products worse.

1.3 What is Opinion Spam

In general, Spam is the presence of any content that is out of place, irrelevant, at times malicious and in general which aims to degrade the authenticity of the digital content in that context. The most common perception of spam is in the form of email spam (where unwanted emails are sent to people's mailboxes) and web spam (where malicious links and content is posted anywhere possible, in forums, chat rooms, blogs etc.)
However, opinion spam is a bit different from the above. Opinion spam is aimed to sway public opinion in favor or against an entity (person, product, service etc.). This kind of spam looks like any other genuine content but is not. Thus, opinion spam studies question the authenticity of the intent of the person, and not the content itself. In other words, the content posted under this category can otherwise be ver genuine looking when read as standalone. This can be understood better using an example.

Suppose a company X sells a product Y. X hosts information about Y on its site and allows people to post their experience and reviews about Y there. Let us assume that X is a famous company and people visit this site to read about product Y and make decisions based on the reviews written here. Let us also assume that Y is a good product on some arbitrary measure. Amongst all the good reviews, there may be some reviews of the following kind:

- people who have a vendetta against company X and post bad reviews about Y so that X faces defame.
- a rival company Z hires people to write bad reviews about Y and promote W, a product similar to Y made by Z.

- reviews who just want to create a ruckus and post random reviews about Y.(These people also fall under the category of Internet trolls)

Let us consider one more scenario. Let Y be a terrible product. Yet X would want to promote it as a good one to get sales. So it hires a group of people and asks them to write good reviews about Y, pretending to have no ties with X whatsoever.

In all the cases above, there is one thing common in the reviews that are being written. The reviews that would be written would all be seemingly genuine. However they do not reflect the true nature of the product Y and hence are deceiving to the person who is reading them. Since we believe people rely on these reviews to make purchase decisions, the reviews are misguiding the customer. This kind of spam is called opinion spam.

1.4 Issues with Spam detection

The above discussion shows how important it becomes to filter out opinion Spam from the reviews. However there are some issues that make detection of this kind of Spam inconvenient:

- The untruthful reviews are crafted to look like genuine reviews. This means that there is no way we can tell if the review is genuinely written or not just by reading these reviews in isolation. They look like any other review.

- Reviews are highly subjective. Most sites enforce minimal to no control over what the person can write as a review. and hence, can vary from a short simple description(even an emoticon) to a long paragraph, describing all pros and cons. All of them would be legitimate reviews.

- There are a number of on-line sites available nowadays. This means that there are a lot of places from where the customer can buy something. This makes it very difficult to ascertain if the reviewer has actually bought and used the product or he is just faking it. Also a universal cross matching of user data from various websites to identify a person is not possible as people use different

aliases in different websites. Moreover, companies cannot share this privileged information anyways.

- Sarcasm and witty reviews are commonplace in the on-line world. Such reviews are tougher to analyze, given that there is no artificial methodology to detect sarcasm correctly, yet. Also sarcasm tends to project a contradictory sentiment to what is actually being written. So, what may appear to be an appreciative review can be a derogatory one and vice-versa.

- Most sites take measures to restrict bots and scripts from writing reviews on products. While this is an advantageous feature, it leads to more difficult problem in the context of opinion Spam detection. A bot generated pattern can be looked for common repetition of words and other such features. Since most review Spam is hand written, a common template for the review text is absent.

- There is no publicly-available tagged dataset for Spam and Ham available to train classification models. Companies hosting these reviews cannot release user information crucial for this analysis in public domain even after making it anonymous. Moreover, human tagging of the reviews is not efficient enough to be used for training model. This has been discussion in Section 2.

1.5 Motivation and Objectives

Given the impact opinion have in the society, it becomes a crucial area of study. In a way, spammed reviews are more of lies than outliers. The proposed work intends to see spam reviews as lies and tries to incorporate the markers used in lie detection in general, in the study.

The main objectives are as follows:

- To study the psycholinguistic features of the reviews in the light of lie detection.

- to simulate simulate markers, used in lie detection in forensics, in text and compare with detection of lies by humans by intuition.

- To use N-gram models to analyze the word sets used by spammers and genuine reviewers.
- To compare and contrast the above methods.

1.6 Thesis organization

This thesis consists of 5 sections, in which Introduction was given in section 1, section 2 gives the literature review on opinion spam. Section 3 gives the proposed framework for study of the same and section 4 gives the results of the study. Section 5 gives the conclusion and future work.

Chapter 2

Literature Review

Despite the gravity of the problem statement, limited work has been done to detect opinion spam. While search engines like Google, Bing etc. invest considerable time and resources in restricting generation of spam content, e-commerce sites have lagged behind in this area. There have been numerous cases reported where individuals and enterprises have admitted spamming the reviews of their own products to promote them in the online market and to create a buzz in the chat forums discussing such items.

Historically, there has been considerable study on Web [2, 6] spam and e-mail [1]. Spam has also been studies in the context of recommender systems[4]. The objective of recommender system attacks is similar to review Spam, their basic ideas are quite different. In recommender systems, a spammer injects some fake profiles in the system to get some products less (or more) frequently recommended.

2.1 Nature of Spam in reviews

Spam in general means any content being posted or propagated that is out of place, irrelevant or malicious in nature. Review spam can be categorized broadly into following categories

1. **Falsified Opinions, also known as untruthful or fake reviews.**
   These reviews have information that do not reflect the true nature of the product being reviewed and are often pushed into the system with a malicious intent. It is generally written with following aims:

- To damage the reputation o some other target product(defaming spam). This type of spam would come under the first category discussed in Section 1.3.
- To promote some product which may or may not be of good nature. This spam is called hype spam and is used bu authors to promote their books in forums, by companies to promote their new released or legacy products etc. The review may attempt to hype up the product to make it look good.

2. **Non-reviews**

   These kind of reviews have little or no information about the product whatsoever. The reviews is basically aimed to divert user attention to something that is in the interest of the spammer-cum-reviewer. The reviews can be one or many of the following:

   - Adverisement about other products that are not generated by the system. The advertisement could be of anything at all(reference to other sites, links to other products, links to other companies or hotels).
   - Link spam is also prevalent in this category, where reviewer post phishing links to deceive the users.
   - reviews with irrelevant information with no opinion about the product.

3. **Reviews on brands only** These reviews do not comment on the products. They make talk about the the manufacturers or the sellers or the brand in general. These have no usefulness in the context of a particular product.

In the above three, Type 1 spams are the most difficult to detect. Type 2 spam which occurs in reviews much less compared to the other two, is non- fatal as human readers detect advertisements pretty well and the spammers intent is not fulfilled here. Type 3 is a comparatively more manageable spam category as the reviews talking about a specific brand is bound to use words from a list of keywords describing the brand. This can be resolved as spam by analyzing the description of the product and that if the user talks only about the company. Usually, human readers are capable in

filtering out reviews in the latter 2 categories as they do not provide info about the product they are looking at.
Thus the primary focus of our analysis to detect Type 1 spam.

2.2 Types of Spammers

Review spammers can be put under one (or more) of the following categories:

1. Individual spammers who just like to create a ruckus.In internet language, they are known as trolls. They just post nuisance for the sake of it and have no economical benefit.

2. Individual spammers who are hired and paid to write reviews.These people write the most genuine-looking fake reviews.

3. individual handling multiple profiles to write spam for greater monetary or psychological benefit.

4. A group of individuals who are hired and paid to write reviews, identified and referred to as *"Spammer groups"*.

A few observations can be made about the nature of spam posted by these reviewers:

- Given that the spammers write reviews professionally, it can be assumed that they have a set of words they may use frequently. Writing a fresh review in itself every time is time-consuming.

- The spammer may want to mix in with the reviews that other people have written for the same product and hence his review can be very similar to the other reviews on the same product.

- The time when a review is posted is crucial. Early reviews get more weightage than the later ones.

- Some sites provide a helpfulness score for the reviews, which can be seen as an indication to the authenticity of the review.

- Most people write reviews after certain intervals, i.e. when they buy something. A spammer is not bound by this instinct. Reviews written in quick succession can be seen as a red flag.

- Since adding any irrelevant data to the review does not help serve the purpose of the spammer, the fake reviews are often crafted as per the sentiment they are trying to portray, making it difficult to detect them.

2.3 Spam detection methods

The current methodology of spam detection can be categorized in two broad categories

1. **Supervised detection methods**

   Supervised learning is the most common form of spam detection in emails. However for reviews, such a tagged data set was not available till late. So limited research has been done in this regard. However, after scraping the websites that host reviews and hiring people to write fake reviews a dataset was prepared for analysis. The details are discussed in the Section 3.1. Supervised techniques work on a simple train-learn-categorize model. The system is provided with a set of training data points. based on this, inferences are drawn regarding the nature of data. These inferences form out supervised classifier as the labels to data points is also available. The trained model is then used against a test data set to check for the validity and accuracy of the model.

2. **Threshold based detection method.** Although the trained classifiers have stronger generalization ability. In practice, the fake reviews occupies a small proportion of all reviews. Thus, it would take some time to gather a certain amount of fake reviews for training. Until that, the fake reviews would make the negative influence. Various parameters can be used with limiting values to indicate that a particular review by a particular user (a change in user for the exact same review would lead to different results) may be spam. These methods are used in starting condition when enough data about spam is not available. The idea is to speculate about certain features of the reviews (for example when

was it written, how correct the language is etc.) and try to associate it with some threshold value to indicate fraudulent behavior

Chapter 3

Proposed Work

3.1 Data Collection

There aqre a very few tagged datasets available in the domain of opinion spam and most work done is based on heuristics. However, such a set can be formed by data from the following sources.

3.1.1 Yelp

Yelp.com is a crowd-sourced website that reviews local businesses. It also has features similar to a social networking site where users can interact with each other as well. Yelp is one of the most popular sites in major metropolitan areas. The site allows users to provide a rating to a product or service, primarily hotels and restaurants, using a one to five star rating. The users can also write descriptive reviews if they wish to. Along with this, Yelp allows its users to check-in into locations they are visiting. The site has 132 million monthly visitors and around half a billion reviews to flaunt. Though Yelp does not provide its dataset publicly, the reviews and user information can be scraped from the site itself as the defense against bots and scripts is kept pretty low to allow more search engine penetration.

3.1.2 Trip Advisor

TripAdvisor is an internet based company that works in travel and tourism business. It works on user-generated content in the form of reviews, about various travel related content. Users can check in using TripAdvisor to various locations they visit. Teh

service is provided free of cost. The site is one of the largest in the world and has more than 60 million members. This large user base has allowed it to garner over 170 million reviews and opinions of hotels, restaurants, attractions and other travel-related businesses making it a good point for study for opinion spam.Similar to Yelp, reviews can be scraped from the site to form review dataset for analysis.

3.1.3 Amazon Mechanical Turk

The Amazon Mechanical Turk is a marketplace which provides personnel on demand. AMT People(called Requesters by the site) are able to post tasks known as HITs (Human Intelligence Tasks). Workers (called Providers or Turkers by the site) can browse the posted tasks and complete them for a monetary payment. This service provides a tool to generate content for research as per our need. The workers were asked to write reviews for hotels specified such theat the reviews were accepted as genuine. This is a convenient way to generate spam for research purpose.

3.1.4 Dataset description

The review dataset made from above sources was compiled as a set of 1600 reviews from 20 hotels in the Chicago region. Each review has the following features:

1. A review ID to uniquely identify each review.
2. Name of the hotel about which the review is written.
3. The review content
4. Review polarity in terms of positive or negative sentiment.
5. a binary tag of the review being spam or not.

This data corpus contains:

- 400 truthful positive reviews from TripAdvisor[8]).
- 400 deceptive positive reviews from Mechanical Turk[8].

- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp[7].
- 400 deceptive negative reviews from Mechanical Turk[7].

There are 20 reviews for each category for each of the 20 most popular Chicago hotels. The hotels under analysis were Affinia, Allegro, Amalfi, Ambassador, Conrad, Fairmont, Fardrock, Hilton, Homewood, Hyatt, Intercontinental, James, Knickerbocker, Monaco, Omni, Palmer, Sharaton, Sofitel, Swissotel and Talbott.
This dataset is particularly useful because of following reasons:

1. The data is well-balanced and has an equal number of reviews for all hotels.
2. The reviews are of both positive and negative polarity and, amongst themselves, contain equal numbers number of spam and ham reviews. This removes the class imbalance problem without any further effort.
3. To get the deceptive reviews, workers at Amazons Mechanical Turk were asked specifically to write a review in such a fashion that it is accepted as a good review by the hosting site.
4. Moreover, the AMT workers had access to the reviews already written about the hotel and could have easily manipulated their entry based on them. This, in turn, simulates the information available to a spammer and his intent.
5. To make sure the genuine reviews were in fact genuine, the reviews were scraped from the original hosting sites and all non-5-star ratings were removed. Also reviews from first time reviewers were eliminated along with reviews that were too short or too long in comparison to other reviews in the dataset.
6. Given that the reviews come from an English-speaking community, chances of use of words from foreign languages is low, making the analysis comparatively easier.

3.2 Psycholinguistic analysis

Hancock et al.(2008)[3]; ; Vrij et al.(2007) [9]; Mihalcea and Strapparava (2009) [5] showed that lies in textual form are correlated with easily computable linguistic features. We have tried to simulate the markers in audio-visual lie detection in text as well. How the sentences are framed, the choice and nature of words and the alienation to person reference can be used possible markers to show if the opinion is deceptive. Based on this, each reviews was analyzed for the following features.

Table 3.1: Psycholinguistic features

Sr. No	Feature	Remarks
1	Average length of Sentences (words)	Number of words per sentence.
2	Average length of words (characters)	Number of characters per word on an average.
3	Number of capitals	Capitals are used to emphasize on opinions.
4	Number of connectors	Connector and, or , however, etc.
5	Number of digits	Number of numbers used in the review.
6	Number of personal references	Use of mine, my, our to establish trust
7	Number of punctuations	Comma, semicolon, colon, exclamations etc.
8	Number of sentences	How large the review is in terms of sentences
9	Number of shortened words	Words like cant, wouldnt etc.
10	Number of time references	Words like yesterday, tomorrow, now etc.
11	Number of words	Word count of the review
12	Polarity	Positive or negative sentiment

Each features tries to capture an element of psychology of the person writing it. It is observed that usually when people write a review, they would try to convince the reader what they believe, may be either to deceive them or genuinely suggest them something. However it is also seen that people who lie tend to try more to build that trust, and in that attempt give it away. We attempt to focus on these markers. For that a list of features were used. Personal references and references to a time frames are often used to show that they were indeed at the hotel. Use of many digits and time can mean higher amount of specific information. Polarity is an important determiner of spam as a review with a polarity that has a high difference from the mean polarity of the rest of the reviews for the same category, is likely to be spam.However we are not considering that feature as the current dataset does not provide us with enough information for that. Lies are often characterized by long sentences, with a number of

connectors. Writing reviews is an informal act, so people often use informal language and punctuations are often ignored. The number of punctuations can therefore be used to model spam nature. People, while writing in a hurry, miss out on the apostrophe in words like can't, don't, shouldn't etc. Unnecessary and overuse of capitalization can be seen as a genuine reviews, as such reviews attract too much attention and usually skimmed over. So a spammer is expected to keep things simple and hidden in plain sight.

3.3 Text Categorization

In contrast to the psycholinguistic strategy discussed above, our text categorization approach to deception detection allows us to model both content and context with n-gram features. The idea is that people who have actually experienced the hotels will use similar kinds of words to review the hotel, simply because the characteristics of the hotel is unchanging. However the spammers would not be compelled to use the same word set.

We assume that spammers are often paid to write fake reviews. Professional spammers would be used to being compensated regularly for writing such reviews. Such frequent practice would mean that the spammer is used to using a particular set of words and expression.

To model this behavior we consider the unigram and bigram feature sets, with the corresponding features(words and expressions) lowercased and unstemmed. From the training set, a dictionary of unigrams and bigrams was maintained. Each review was then broken into corresponding N-gram and was checked for the following scores. This score is calculated on the basis of presence or absence of an N-gram in the spam set or the ham set in terms of 1 or 0.

Thus, after the analysis the scores would give us an indication of how much the review is similar(and different) to the spam reviews and how much to the genuine or ham reviews.Such score is then used to model the spam behavior. The scores' description is in Table 3.2

Table 3.2: Text categorization features

Sr. No	Score	Remarks
1	Spam Hit Score	No. of N-grams in test review found in Spam N-gram dictionary
2	Ham Hit Score	No. of N-grams in test review found in Ham N-gram dictionary
3	Spam Miss Score	No. of N-grams in test review not found in Spam N-gram dictionary
4	Ham Miss Score	No. of N-grams in test review not found in Ham N-gram dictionary

3.4 Classifiers

Classification is essentially putting a label to a given observation based on previous observations and learnings. In machine learning, classification is the problem of identifying the set of categories a new observation belongs to. The decision is based on training set of data containing observations whose category membership is known. We are using a tagged dataset for spam and genuine reviews. The feature set derived from the above two methods were used to train the following classifiers and then tested against a test set. The observation parameter of comparison was the Accuracy obtained in each feature-classifier pair.

3.4.1 Classification Tree

Classification or decision trees are used in classification models as a series of if-then statements. The process forms a tree like structure where the leaves contain the class labels and the branches show the conjunction of the decisions on features leading up to that class decision. At each internal node, it is attempted to make the data points in one node more homogeneous by partitioning them into two or more sets based on some features which facilitates the division the most. This process is continued till all data points in a node are of the same class and no further division is possible/required. These nodes then become class labeled leaves.

A decision tree is often characterized by an accompanying information gain measure which attempts to model how different the data points are with respect to individual or group of features. In our setup, Gini index was used for that information.

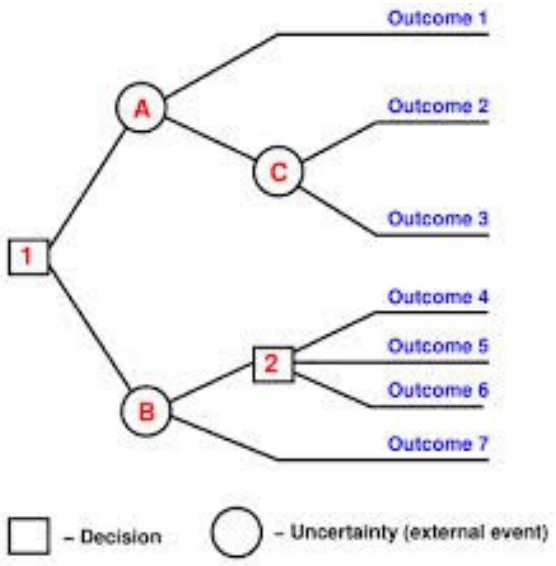

Figure 3.1: Classification tree demonstration

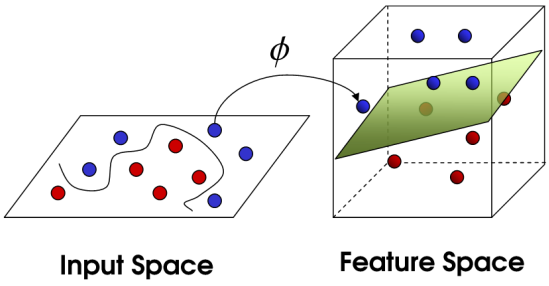

Figure 3.2: Support Vector Machine demonstration

3.4.2 Support Vector Machines

A Support Vector Machine classifies data points in an N-dimensional space by trying to find an N-dimensional hyperplanes which separates the dataa points of different classes. The hyperplane chosen is the one which has the largest margin of differentiation between the given classes.

$$y = \text{sign}(w \quad x + b)$$

In mathematical terms, when a data point is given by a vector of N features, differentiation between these points may not be possible to find, in case a linear demarcation is being looked for. However when projected to a higher dimensional

space, such a demarcation is possible. A general equation for the same can be given as

Chapter 4

Results and observations

4.1 Psycholinguistic analysis

Table 4.1: Psycholinguistic features result

Approach	Features	Train set size (in %)	Classifier	Accuracy
Psycholinguistic features	Linguistic features vector	70	SVM	53.33
			Decision Tree	57.65
		80	SVM	55.34
			Decision Tree	60.11
		90	SVM	52.89
			Decision Tree	56.89

The psycholinguistic model works averagely well and the results are shown in Table 4.1. However an important observation is that such a simplified analysis can also yield results that are comparable to that of a human classifying the same data. Ott et al [8] found that humans has a accuracy of ¡60% for the same dataset. Also, even when multiple volunteers were asked to label this data, the concurrence ate among them was fairly low. Thus, the psycholinguistic model matches the intuition of a human in spam classification.

4.2 Text categorization analysis

Table 4.2: Text Categorization result

Approach	Features	Train set size (in %)	Classifier	Accuracy
Text Categorization	Unigram	70	SVM	76.87
			Decision Tree	92.00
		80	SVM	83.12
			Decision Tree	94.06
		90	SVM	80.62
			Decision Tree	91.25
	Bigram	70	SVM	73.90
			Decision Tree	96.00
		80	SVM	82.18
			Decision Tree	96.87
		90	SVM	80.62
			Decision Tree	97.5
	Unigram + Bigram	70	SVM	67.91
			Decision Tree	95.83
		80	SVM	70.00
			Decision Tree	96.56
		90	SVM	70.62
			Decision Tree	95.00

The results for N-gram text categorization is shown in Table 4.2. The accuracy is better than the psycholinguistic model alone. Following observations can be made about the same:

- The frequent word set used by the spammers and those who write genuine reviews is different enough to help us tag spam behavior. This validates our initial hypothesis.
- Though there are words that are common to both Spam and Ham word sets, the frequency of usage of such words is also a important point to consider.
- The N-gram model can in general be applied to scenarios other than hotels as the basic ideology remains the same.

Table 4.3: Psycholinguistic plus Text features result

Approach	Features	Train set size (in %)	Classifier	Accuracy
Psycholinguistic + Text Categorization	Linguistic features vector combined with Unigram and Bigram values	70	SVM	67.88
			Decision Tree	82.44
		80	SVM	68.97
			Decision Tree	84.67
		90	SVM	66.45
			Decision Tree	83.45

4.3 Combined analysis

The Ngram model alone seems to overfit the data points and does not include any features os the spammer other than the words they use. Combining the previous two models gives more reasonable results and a more realistic modeling of the data set, as shown in Table 4.3. The accuracy levels still remain fairly higher than most work in this area.

Chapter 5

Conclusions and Future Work

The above analysis implies that unigram and bigram analysis work quite effectively in detecting spam based on just the review text. The linguistic features offer secondary support to the decision model. The combined model gives more reasonable results as it also encompasses the psychological tendency of the spammer.

Furthermore, we believe that the spam analysis would be much more powerful if the information about the users is also available. User metadata like number of reviews written, the timeframe in which he writes the reviews, the geolocation or check in data if available from other sources to verify if the user was actually present at the venue, age of the user etc. can be crucial elements in determining if a review is fraudulent. Unfortunately, user information is not divulged to public due to privacy reasons and can only be analyzed internally by the websites themselves.

One other way to improve the spam detection rate is to form a pseudo-truth value for various attributes about the product being reviewed. For instance, while considering electronics goods, reviews from official technology reviewers like Digit, Chip etc. can be obtained to form a document about the item, against which all further reviews can be checked. In the case of hotels, critical reviews from official hotel review boards can be used as the pseudo truth value.

Nevertheless, the proposed work can be used as a base work on which further improvements can be made.

Bibliography

[1] Harris Drucker, S Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054, 1999.

[2] Zoltan Gyongyi, Pavel Berkhin, Hector Garcia-Molina, and Jan Pedersen. Link spam detection based on mass estimation. In *Proceedings of the 32nd international conference on Very large data bases*, pages 439–450. VLDB Endowment, 2006.

[3] Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2007.

[4] Bhaskar Mehta, Thomas Hofmann, and Peter Fankhauser. Lies and propaganda: detecting spam users in collaborative filtering. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 14–21. ACM, 2007.

[5] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics, 2009.

[6] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 83–92, New York, NY, USA, 2006. ACM.

[7] Myle Ott, Claire Cardie, and Jeffrey T. Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Short Papers, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[8] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

[9] Aldert Vrij, Samantha Mann, Susanne Kristen, and Ronald P Fisher. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior*, 31(5):499, 2007.